AFFECTIVE MEETING VIDEO ANALYSIS

Alejandro Jaimes[*], Takeshi Nagamine[*], Jianyi Liu[*], Kengo Omura[*], and Nicu Sebe[&]
[*]*FXPAL Japan, Fuji Xerox Co., Ltd.,* [&]*University of Amsterdam*
{alex.jaimes, takeshi.nagamine, kengo.omura}@fujixerox.co.jp

ABSTRACT

In this paper we examine the affective content of meeting videos. First we asked five subjects to manually label three meeting videos using continuous response measurement (continuous-scale labeling in real-time) for *energy* and *valence* (the two dimensions of the human affect space). Then we automatically extracted audio-visual features to characterize the affective content of the videos. We compare the results of manual labeling and low-level automatic audio-visual feature extraction. Our analysis yields promising results, which suggest that affective meeting video analysis can lead to very interesting observations useful for automatic indexing.

1. INTRODUCTION

Many researchers are working on developing techniques to automatically index meeting video contents (e.g., speech recognition, action recognition, etc.). In spite of these efforts, little attention has been paid to the affective content of meeting videos, although it has been an area of interest for many years in psychology and communications research.

Automatically indexing the affective content of meeting videos is important for several reasons. First, research has shown that there is a strong relationship between affective content and memory: we are more likely to remember events associated with strong emotions (i.e., indexing strong emotions is useful). Second, affective content indexing adds value to the videos (i.e., non-obvious observations added). Finally, it can contribute to research in psychology and communications projects that require manual labeling.

In this paper, we investigate the affective content of meeting videos and its relation to low-level automatic audio-visual feature extraction. First, we performed an experiment in which 5 subjects labeled 3 real meeting videos of 10 minutes each to judge each meeting's *valence* (pleasure/displeasure) and *arousal* (excited/calm) contents using *Continuous Response Measurement (CRM)*. Then, we automatically extracted audio-visual features to measure valence and arousal. We compare manual and automatic results in order to answer the following three questions:

- What is the affective content of meetings like?
- Can CRM be used to measure the affective content of meetings and are such measures useful?
- How do CRM measurements relate to automatic analysis (can we approximate CRM responses using audio-visual analysis)?

Our work differs from related research in the following aspects: (1) CRM has not been applied to meeting videos before (in contrast to labeling using descriptive text [13], interval-coding to obtain binary labels [5], and labeling of utterances [9]), and (2) we investigate the span of the affective space occupied by meeting videos (in contrast to finding *only* hot spots [14][4]). In addition, we apply our approach to real (not mockup) meetings (unlike [5][13]), and examine how low-level audio-visual features correlate with CRM.

The rest of the paper is organized as follows. In section 2 we discuss CRM and affective space. In sections 3 and 4 we describe our audio-visual analysis approach. We discuss experiments in section 5 and conclude in section 6.

2. CRM & THE AFFECTIVE SPACE

2.1.Continuous Response Measurement

Common techniques to study human behavior include sequential analysis [1] and CRM [2]. In the former, behavior patterns are quantitatively studied (e.g., for conflict resolution, police investigations, learning, decision-making processes, etc.) by labeling communicative acts using pre-defined categories. In CRM, a subject (e.g., TV viewer), uses an input device (e.g., a dial with a continuous scale) to "self-report" changes in a psychological state or judgement, changes in message content, or to code communication behaviors, in real time. For example, while watching a video, a subject moves a dial up or down to indicate that he "likes" or "dislikes" what he is viewing. CRM has been used widely to study patterns of interpersonal communication, cognitive responses to speeches, TV programs, ads, and others, since as early as 1945 [2],

but to the best of our knowledge, it has not been applied to meeting analysis (see [12] for other related research).

2.2. Affective Space

Human affective response is usually represented using three dimensions [6]: *valence*, *arousal*, and *control* (dominance). Valence (i.e., type of emotion) is typically characterized as a continuous range of affective responses or states extending from pleasant or "positive," to unpleasant or "negative." Arousal (i.e., intensity of emotion) is characterized by affective states ranging on a continuous scale from energized or alert, to calm or peaceful. Typically, arousal and valence values are plotted together to generate a 2D affect curve, to which any human emotion can be mapped (figure 6). The control dimension is generally neglected because it has been shown that it plays a limited role in characterizing various emotional states [6].

We have chosen CRM to obtain the manual valence and arousal curves because it has been used to measure affective response, as well as perceptual and semantic judgments and processes [2]. It has also been used to measure attention and shown to predict memory of particular video segments. Another alternative (used for meeting videos in [5]) is interval coding, in which labels are assigned every *t* seconds. Events, however, may occur between intervals, it is difficult to determine the right interval length size, and labeling is not done in real time. CRM does not have these problems (see [2] for a detailed discussion on advantages of CRM).

3. VISUAL ANALYSIS: AROUSAL

Psychologists, social scientists, and anthropologists have long been interested in analyzing visual, non-verbal communication because it conveys important affective information. The "Adam's-apple-jump", for example, is described as an unconscious sign of emotional anxiety, embarrassment, or stress [11]. At a business meeting, a listener's Adam's apple may inadvertently jump if he strongly disagrees with a speaker's suggestion, perspective, or point of view. The conference table is described as a "nonverbal battlefield," where to promote key points, speakers lean forward over the table and use palm-down gestures.

Visual information can clearly contribute to identifying interesting affective meeting content. In [7] we used a simple activity measure based on frame differencing, skin, and face detection to automatically structure multi-stream meeting videos for browsing (but we did not compare the results with continuous labeling or consider affective content as we do here). Although the activity in [7] and arousal are two different concepts, the measure produces interesting results. Figure 1 shows the meaning of some of the activity peaks. A comparison of automatic and manual results is shown in Figure 5 (after scaling for comparison).

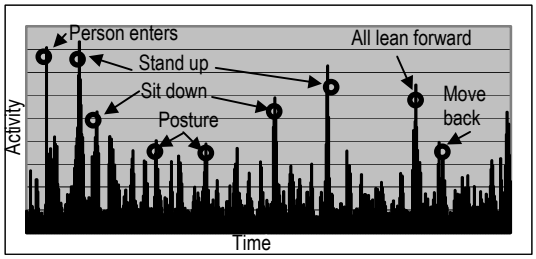

Figure 1. Global activity measure obtained automatically for one of the videos (video A).

We also applied the Visual Trigger Templates framework [8] to detect large posture changes, which may indicate interest level changes [10]. The templates are manually constructed only once for each video and the analysis is done automatically. A template consists of a set of bounding boxes and template trigger conditions (e.g., a template box is triggered when there is motion inside the box). The dotted box in each image of Figure 2, for example, indicates that the template is "triggered"—the person of interest is inside the template. This way, local posture changes can be easily measured to detect individual activity.

Figure 2. Posture changes and a template for a particular individual.

Although not all participants make drastic posture changes, we have found that in many cases several attendees make similar body motions concurrently (Figure 2, right). This often coincides with a high arousal rating by the human observers, and show up as a peaks in the automatic activity measure (Figure 1).

4. AUDIO ANALYSIS: VALENCE

Automatic pitch estimation has been used to detect changes in emphasis in spoken dialogue (e.g., [9][14]), and a correlation between pitch and valance has been shown experimentally [6]. Since pitch values can differ significantly for different individuals, it is desirable to estimate each person's pitch distribution. Unfortunately, this requires training, and high-quality (e.g., neck) microphones for automatic speaker separation.

Our meeting room has standard tabletop microphones, which make automatic speaker separation difficult. However, the pitch distributions are similar in the videos we are using (all of the attendees are Japanese males). Therefore, we assumed a single pitch

distribution for all speakers. Although we found that this assumption works reasonably well, speaker separation is needed if there are wide variations in the attendees' pitch distributions.

We extract the following features from the audio signal: average pitch, maximum pitch, and pitch standard deviation. We used the MAD framework in Matlab to extract and analyze the pitch signals [3]. Pitch was obtained using frames of length 1024 (32 kHz sampling rate) and one second segments. Silence segments were eliminated by counting the number of non-zero pitch values within each segment (if there are fewer than 6 non-zero pitch values, the segment is silent). This simple method, although very restricted, yielded reasonable results for our meetings since silence periods in meetings are characterized by white noise (very low pitch).

5. EXPERIMENTS

5.1. Experimental Setup

We randomly selected the first ten minutes of 3 real research discussion meetings (each with 4-5 different male attendees, held about 1 month apart, independently of this study). The meetings were held in Japanese by native Japanese speakers. Then, we asked 5 native Japanese speakers (one female and four males in their 20s and 30s, who did not participate in the meetings), to manually label each of the videos using a GUI we constructed for continuous response measurement (Figure 3). The subjects could move a dial in real time while observing their response curve. The motivation for this is that the response level at time t is more easily determined by the subject if he is able to view the values he has assigned so far (start of video, t_0 to current time t),

Figure 3. Our CRM interface. The subject views the video and the curve depicting the values he has assigned so far.

The subjects were instructed to label the videos using only the dial based on what they thought were the *meeting participants'* arousal ("level of excitement") and valence levels ("measure of positive or negative feelings"). Each meeting was viewed twice by each subject: once for arousal and once for valence. The subjects were not familiar with the automatic analysis approach used in this paper and they were not told how the data would be used. They were instructed to pause the video only in case of error.

5.2. Results

For each video (A, B, C) we obtained 10 curves from 5 subjects: 5 for valance and 5 for arousal. We extracted the visual activity features described in section 3, and the pitch features described in section 4.

We found the results of manual labeling quite interesting, particularly those of meeting A, shown in Figure 4 (we marked some of the big changes in the average curves). As expected, there were variations in the labeling across the 5 subjects (one of the curves is barely visible because it is so close to the "neutral" line), but there are clear similarities between the curves.

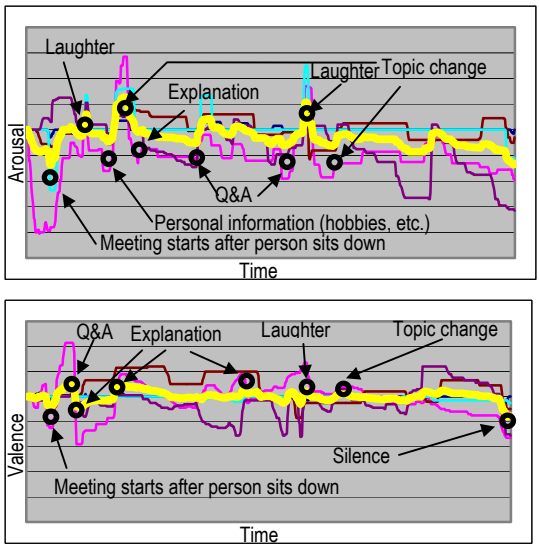

Figure 4. Arousal (top) and valence (bottom) labeling by each subject for video A. The thick line represents the average for all three in each case.

In that meeting, a new summer student was asked to introduce himself and discuss his research. The meeting starts with a low arousal level as people get ready for the discussion. The student introduces himself (since some attendees have not met him) by speaking of his hobbies and personal history (a common pattern in Japanese meetings when someone is first introduced), which brings up the arousal and valence levels. Since it is the student's first meeting, he is nervous and hesitates in his explanations when questions are asked, thus there are drops in the valence curve at several explanation periods. Valence goes up with laughter.

Figure 5 shows the manual and automatic results for arousal and valence. For discussion purposes, we plot only the global visual activity curve for video A (similar results are obtained detecting local posture changes), and a time-aligned, low-passed filtered (window size 15 seconds) version of the average pitch values for video B, whose audio results are more interesting than those of video A (although similar).

The similarity in the general trends of the curves is not surprising and most visible in some sections when affect changes manifest themselves audio-visually (compare with Figure 4). People lean forward when they are looking at something together during the meeting, or when group interest raises, and audio pitch values change as the mood in the meeting changes. As a simple quantitative measure, we thresholded both curves to obtain binary labels and compared retrieval of one vs. the other. We obtained 85% accuracy for valence (11% miss and 4% false alarm) for video B and 81% accuracy for arousal (4% miss and 15% false alarm) for video A. Thresholds were chosen empirically and other results were similar.

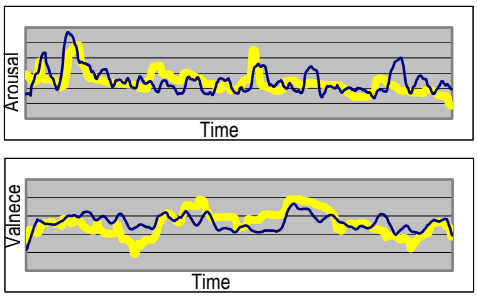

Figure 5. Manual average arousal curve and automatic visual activity curve (top, for video A). Manual average valence curve and automatic *mean* pitch curve (bottom, for video B). Manual curves are thick yellow. Curves were smoothed, scaled, and shifted to account for delay errors.

We also mapped the two dimensions (valence and arousal) to the 2D affect curve in Figure 6, for video A. The purpose of this is to show that, over time, the manual labels span a fairly wide range of the affective space as a result in changes in meeting participants' activity (arousal), and interest (valence).

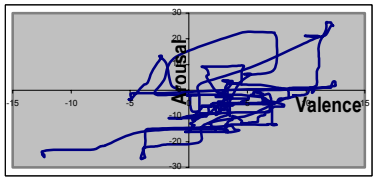

Figure 6. Affective space mapping for video A, using the average scores of all 5 subjects.

6. CONCLUSIONS & FUTURE WORK

We examined the affective content of meetings. We asked 5 users to manually label 3 meeting videos in real time using continuous scales for arousal and valence (continuous response measurement). Then we automatically extracted audio-visual features to characterize the affective content of the meetings and compared the results of manual and automatic labeling.

Although we used simple automatic techniques that could be greatly improved, these early results are very promising. First, we found that continuous response measurement (CRM) is useful: it shows that the affective content of meetings can vary significantly, and that obtaining affective information can yield interesting results. Second, we found that low-level features, in particular motion activity (visual) and pitch (audio) can produce curves that have important similarities to curves generated manually by observers using CRM. The results, however, should be interpreted with caution, as more detailed analysis is required.

Future work includes further analysis, a larger study, extraction of additional audio-visual features, exploring the use of CRM during the meeting (e.g., annotations for audio recordings in [4]), and retrieval experiments.

7. REFERENCES

[1] R. Bakeman and J.M. Gottman. *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge, 1997.

[2] F. Biocca, P. David, and M. West, "Continuous Response Measurement (CRM): A Computerized Tool for Research on the Cognitive Processing of Communication Messages," in A.Lang, (ed.), *Measuring Psychological Responses to Media Messages*, L. E. Associates, NJ, 1994.

[3] M.P. Cooke, and G.J. Brown, "Interactive explorations in speech and hearing." *J. Acoust. Soc. Japan (E)*, 20, 2, 89-97, 1999 (http://www.dcs.shef.ac.uk/~martin/).

[4] N. Eagle and A. Pentland, "Social network computing," In *Proc. UBICOMP*, Seattle, Oct. 2003.

[5] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting Group Interest-Level in Meetings," *IDIAP Research Report 04-51*, September 2004.

[6] A. Hanjalic. *Content-Based Analysis Of Digital Video*, Kluwer, 2004.

[7] A. Jaimes, et. al., "Interactive Visualization of Multi-Stream Meeting Videos Based on Automatic Visual Content Analysis" in *IEEE MMSP '04*, Siena, Italy, Sept. 2004.

[8] A. Jaimes, Q. Wang, N. Kato, H. Ikeda, and J. Miyazaki, "Visual Trigger Templates for Knowledge-Based Indexing," in *IEEE PCM 2004*, Tokyo, Japan, Dec. 2004.

[9] L. Kennedy and D. Ellis, "Pitch-based Emphasis Detection for Characterization of Meeting Recordings," in *Proc. ASRU 2003*, Virgin Islands, Dec. 2003.

[10] S. Mota and R. Picard,"Automated posture analysis for detecting learner's interest level," In *Proc. CVPR Workshop on CVPR for HCI*, Madison, Jun. 2003.

[11] Non-verbal dict.: http://members.aol.com/nonverbal2/diction1.htm#The%20NONVERBAL%20DICTIONARY

[12] A. Pentland, "Socially Aware Computation and Communication," *IEEE Computer,* Vol.38:3, March 2005.

[13] P. Wellner, M. Flynn, S. Tucker, and S. Whittaker, "A Meeting Browser Evaluation Test," *IDIAP RR-04-53*, 2004.

[14] B. Wrede and E. Shriberg, "Spotting Hotspots in Meetings: Human Judgments and Prosodic Cues," In *Proc. Eurospeech*, Geneva, Sep. 2003.